Regression with Gaussian Process Networks

Manel Martínez-Ramón

ECE, UNM

October, 2018

Regression in Hilbert spaces

- The solution of the dual parameters is $\boldsymbol{\alpha} = \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y}$ where

$$\mathbf{K} = \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_p \boldsymbol{\Phi}$$

  which defines kernel

$$k(\mathbf{x}, \mathbf{z}) = \boldsymbol{\varphi}(\mathbf{x})^\top \boldsymbol{\Sigma}_p \boldsymbol{\varphi}(\mathbf{z}) = \boldsymbol{\varphi}(\mathbf{x})^\top \boldsymbol{\Sigma}_p^{1/2} \boldsymbol{\Sigma}_p^{1/2} \boldsymbol{\varphi}(\mathbf{z})$$

  which is a dot product of vectors linearly transformed by matrix $\boldsymbol{\Sigma}_p^{1/2}$.
- The choice if this matrix is implicit in the choice of the kernel.

- With the previous definition of the kernel, the regression can be written as

$$
\begin{aligned}
\bar{f}_* &= \boldsymbol{\varphi}(x^*)^\top \bar{\mathbf{w}}' \\
&= \boldsymbol{\varphi}(x^*)^\top \boldsymbol{\Sigma}_p \boldsymbol{\Phi} \boldsymbol{\alpha} \\
&= \mathbf{k}^\top(\mathbf{x}^*) \boldsymbol{\alpha} \\
&= \mathbf{k}^\top(\mathbf{x}^*) \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y}
\end{aligned}
$$

where $\mathbf{k}^\top(\mathbf{x}^*) = \{k(\mathbf{x}^*, \mathbf{x}[1]) \cdots k(\mathbf{x}^*, \mathbf{x}[N])\}$ is the column vector of dot products between the test data $\mathbf{x}^*$ and all training data $\mathbf{x}[n]$ into the Hilbert space.

- Expression

$$\bar{f}_* = \mathbf{k}^\top(\mathbf{x}^*)\left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1}\mathbf{y}$$

  is the *expectation of the Gaussian process* at sample $\mathbf{x}^*$ conditional to $\mathbf{X}, \mathbf{y}$.

- We can compute the variance of the process at this point simply using the obtained definition of the kernel. Indeed:

$$\sigma_{f_*}^2 = \boldsymbol{\varphi}(\mathbf{x})^{*\top}\mathbf{A}^{-1}\boldsymbol{\varphi}(\mathbf{x}^*)$$

  with $\mathbf{A} = \sigma_n^{-2}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \boldsymbol{\Sigma}_p^{-1}$.

- We compute the inverse of $\mathbf{A}$ using the matrix inversion lemma:

$$(\mathbf{UWV}+\mathbf{Z})^{-1}=\mathbf{Z}^{-1}-\mathbf{Z}^{-1}\mathbf{U}(\mathbf{V}^\top\mathbf{Z}^{-1}\mathbf{U}+\mathbf{W}^{-1})^{-1}\mathbf{V}^\top\mathbf{Z}^{-1}$$

Since $\mathbf{A}=\sigma_n^{-2}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top+\boldsymbol{\Sigma}_p^{-1}$ we can identify

$$\mathbf{Z}=\boldsymbol{\Sigma}_p^{-1},\quad \mathbf{U}=\mathbf{V}=\boldsymbol{\Phi},\quad \mathbf{W}=\sigma_m^{-2}\mathbf{I}$$

from which

$$\mathbf{A}^{-1}=\boldsymbol{\Sigma}_p-\boldsymbol{\Sigma}_p\boldsymbol{\Phi}\left(\mathbf{K}+\sigma_n^2\mathbf{I}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}_p$$

- Now, from expression

$$\sigma^2_{f_*} = \boldsymbol{\varphi}(\mathbf{x})^{*\top} \mathbf{A}^{-1} \boldsymbol{\varphi}(\mathbf{x}^*)$$

we obtain

$$\begin{aligned}\sigma^2_{f_*} &= \boldsymbol{\varphi}(\mathbf{x})^{*\top} \left(\boldsymbol{\Sigma}_p - \boldsymbol{\Sigma}_p \boldsymbol{\Phi} \left(\mathbf{K} + \sigma_n^2 \mathbf{I} \right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_p \right) \boldsymbol{\varphi}(\mathbf{x}^*) \\ &= \boldsymbol{\varphi}(\mathbf{x})^{*\top} \boldsymbol{\Sigma}_p \boldsymbol{\varphi}(\mathbf{x}^*) - \boldsymbol{\varphi}(\mathbf{x})^{*\top} \boldsymbol{\Sigma}_p \boldsymbol{\Phi} \left(\mathbf{K} + \sigma_n^2 \mathbf{I} \right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_p \boldsymbol{\varphi}(\mathbf{x}^*)\end{aligned}$$

- Using the definition of the kernel, the variance is

$$\sigma^2_{f_*} = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I} \right)^{-1} \mathbf{k}(\mathbf{x}^*)$$

- The predictive mean and variance of the Gaussian process in a Hilbert space defined by kernel $k(\mathbf{x}, \mathbf{x})$ evaluated at $\mathbf{x}^*$ are

$$\bar{f}_* = \mathbf{k}(\mathbf{x}^*) \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y}$$

and

$$\sigma_{f_*}^2 = k(\mathbf{x}*, \mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{k}(\mathbf{x}^*)$$

where $\mathbf{k}(\mathbf{x}^* = \{k(\mathbf{x}^*, \mathbf{x}[1] \cdots k(\mathbf{x}^*, \mathbf{x}[N])\}^\top$.

Interpretation of the kernel as a covariance

Proposition: *The kernel matrix* $\mathbf{K}$ *is the covariance matrix of the Gaussian process estimation of* $f(\mathbf{x}[n])$.

Proof: We assume that the process has zero mean, this is $\mathbb{E}[f(\mathbf{x}[n])] = 0$. Then, straightforwardly, the covariance matrix is

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}[n])f(\mathbf{x}[m])] &= \mathbb{E}[\boldsymbol{\varphi}(\mathbf{x}[n])^\top \mathbf{w}'\mathbf{w}'^\top \boldsymbol{\varphi}(\mathbf{x}[n])] \\ &= \boldsymbol{\varphi}^\top(\mathbf{x}[n])\mathbb{E}[\mathbf{w}'\mathbf{w}'^\top]\boldsymbol{\varphi}(\mathbf{x}[m]) \\ &= \boldsymbol{\varphi}^\top(\mathbf{x}[n])\boldsymbol{\Sigma}_p\boldsymbol{\varphi}(\mathbf{x}[m]) = k(\mathbf{x}[n], \mathbf{x}[m])\end{aligned}$$

Corollary There is an error $\varepsilon[n]$ in the estimation

$$y[n] = f(\mathbf{x}[n]) + \varepsilon[n]$$

which is modelled as AWGN, independent of $\mathbf{x}[n]$ and with variance σ_n^2. Then, the covariance of the regressors is

$$\mathbb{E}[y[n]y[m])] = k(\mathbf{x}[n], \mathbf{x}[m]) + \sigma_n^2 \delta(m - n)$$

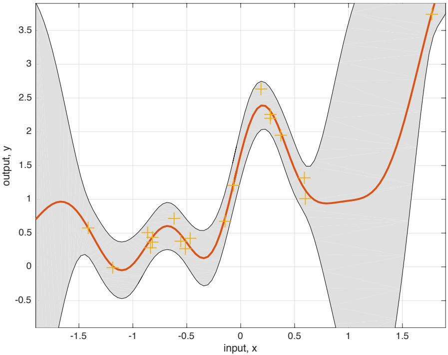
so the covariance matrix of the process is

$$\mathbf{C}_{yy} = \mathbf{K} + \sigma_n^2 \mathbf{I}$$

- The previous result gives an alternative interpretation of the predictive likelihood with the mean and variance summarized in slide 11.
- Assuming a set of training data $\mathbf{X}$ and corresponding regressors $\mathbf{y}$, and a test data $\mathbf{x}^*$, the joint process $\mathbf{y}, f_*$ is, from the previous results

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma_n^2 \mathbf{I} & \mathbf{k}(\mathbf{x}^*) \\ \mathbf{k}^\top(\mathbf{x}^*) & k(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix}\right)$$

Using the Bayes rule, one can compute the pdf of f_* conditional to $\mathbf{X}, \mathbf{y}$, which is a Gaussian with the mean and variance of slide 7.



Example included in the software provided in Rasmussen et al, 2006. A set of samples is generated from a filtered Gaussian distribution. The line represents the predictive mean. The band represents the standard deviation of the prediction.

Outcomes of this lesson

After this video, students should be able to:

- Identify the mean and variance of the predictive posterior in a Kernel Gaussian Process
- Reproduce the proofs for both expressions.
- Prove that the kernel matrix is equal to the covariance matrix of the training regressors.